ESPC and NUOPC Workshop on Construction, Visualization, and Verification of Ensemble-based Predictions from Sub-seasonal to ISI Timescales

30 July-1 August 2013 at the Scripps Institute of Oceanography

Participants:

Stan Benjamin	NOAA/ESRL
Michal Branicki	NYU
Barbara Brown	NCAR/RAL
Jessie Carman	NOAA (ESPC DPM)
David Considine (Remote)	NASA
Dan Copsey	UK MET Office
Bruce Cornuelle	SIO
John Cortinas	NOAA OAR
Tim DelSole	GMU
Huug van den Dool	NOAA CPC
Dan Eleuterio	ONR (ESPC PM)
Orlando Florez	Navy ONR
Josh Hacker	NCAR/RAL
Tom Hamill	NOAA/ESRL/PSD
Jim Hansen	Navy NRL MRY
Scott Harper	Navy ONR
Pat Hogan	Navy NRL SSC
Jin Huang	NOAA CPC
Gregg Jacobs (Remote)	Navy NRL SSC
Ben Kirtman	U. of Miami/RSMAS
Kurt Lutz (Remote)	NUOPC Staff
David Margolin	EarthRisk Technologies
Julie McClean	SIO
Justin McLay	Navy NRL MRY
Art Miller	SIO
Edward C. Mozley	Navy PEO C4I PMW 120
Carolyn Reynolds	Navy NRL MRY
Scott Sandgathe	UW (ESPC Staff)
Liz Satterfield	Navy NRL MRY
Gabe Vecchi	NOAA GFDL
Steve Warren	NOAA (ESPC Staff)
Tim Whitcomb	Navy NRL MRY

Summary:

An ESPC and NUOPC Workshop on Construction, Visualization, and Verification of Ensemble-based Predictions from Sub-seasonal to ISI Timescales was held on 30 July-1 August 2013 at the Scripps Institute of Oceanography in San Diego. The agenda for the workshop was:

- Welcome/Overview
- Session 1: Climate Verification vs. Weather Verification - Looking at Things Differently
- Session 2: Verifying Multi-Model Coupled Systems - How Do You Trace Errors and What Metric Do You Use?
- Session 3: Systematically Accounting For Model Inadequacy

- Session 4: Post Processing and Visualization on ISI Timescales
- Recap/Development of Recommendations

Dan Eleuterio, the ESPC Program Manager, and Cathy Constable from the Scripps Institute of Oceanography welcomed workshop participants. The workshop goals were presented as:

- This workshop will consider science issues governing the optimal use and configuration of single and multi-model global coupled earth system ensembles (MMEs) for sub-seasonal (synoptic) to intra-seasonal and inter-annual (ISI) timescales and the simulation, quantification and presentation of forecast uncertainty
- The workshop format will emphasize discussion with the goal of elucidating science and technical challenges to ISI prediction and opportunities for achieving useful products

The first session focused on "Climate Verification vs. Weather Verification." During the session, Barbara Brown discussed thoughts on verification at ISI timescales. She spoke of the concept of seamless verification across space/timescales and posed the question of whether skill can be shown at the longer timescales. She also questioned whether seamless verification was practical – greater aggregation may lead to reduced specificity. Additionally, it was noted that problems and decisions are different on different timescales. The same evaluation approaches may not be appropriate for different timescales. Unique but overlapping verification methods for the timescales may be appropriate. The need for identifying the most meaningful metrics for different timescales was discussed along with the need to determine how to provide representative information that represents the needs of a range of users. Both predictability and limits on availability of information were believed to be factors limiting skill at longer timescales. A suggestion was made to increase the use of skill verifications for patterns of phenomena. This is currently done mostly in a subjective manner and should be made more objective. There is also a need to increase specificity on forecasts at longer timescales which in turn would allow for more objective/specific verification (for example for longer timescales, it would be useful to know not only that it will be colder/warmer than normal but also how much colder/warmer than normal).

Tom Hamill briefed "Blocking and the Madden Julian Oscillation (MJO) in GEFS Reforecasts: Forecast Skill and Interactions." He noted a desire to find better ways of visualizing verifications. Use of reforecasts to examine the ability of the forecast model to skillfully predict low-frequency modes of variability and longer-lead forecasts was also mentioned. Skill has been shown for blocking out to 13 days. Teleconnections between MJO's and features in other ocean regions were shown. These were noted as complex connections not handled well by current models. It was noted that connections involve interactions through the stratosphere and increased vertical resolution in the models would be helpful in improving modeling skill. Emphasis to date has been more on horizontal rather than vertical resolution. Tom Hamill's conclusions were:

- Blocking:
  - Some skill, but much less than a perfect model prediction. For intraseasonal forecasts, little skill exists by week +2
  - Reasonable replication of blocking climatological frequencies in forecasts
- MJO:
  - Forecasts decrease in MJO amplitude, slow down progression relative to analyzed
  - Ensemble forecasts under-dispersed/biased, especially for magnitude
  - Some skill, though, especially for high amplitude MJOs
- Blocking and MJO:
  - Blocking frequency changes in response to active MJO not correctly forecast except with shorter-term forecasts

Tim DelSole briefed “A Rigorous Framework for Validating Ensemble Forecasts.” He posed questions of how forecast skill/scores are best compared and how big of a difference in score is meaningful. He summarized that:

- Testing the significance of a difference in skill is difficult because most skill metrics use observational estimates. The fact that the same observations are used implies that the skill metrics are not independent.
- The dependence in the skill metrics can be overcome in well calibrated systems by comparing forecast spread
- For the National Multi-Model Ensemble:
    - Mean Square Error (MSE) intervals are large and overlap with each other
    - MSE is consistent with forecast spread for 4 models of the multi-model and of these, multi-model forecast has significantly worse score
    - Every model has periods in which the multi-model enhances skill
    - The multi-model systematically improves skill during spring barrier

In the Session 1 discussion period, several overarching questions were posed. For ISI timescales, do we emphasize climate techniques or some combination or variation in order to achieve the best results – for diagnostics? For user metrics? What are the correlations between a model’s weather forecast skill and climate forecast skill? Which aspects of weather simulation are important for ISI validation? This gets back to that old question of whether it is beneficial for a weather model to be a good climate model. Can we succeed in developing decision relevant climate validation? From the discussion of these questions, a recommendation was made to develop an aggregate map of user needs/decision maker requirements. This would likely require a sector by sector review (different user communities such as industry). Related to this, we need visualization of what we can predict at various timescales. A recommendation from this would be to develop a global visual depiction of predictability skill for different forecast values or phenomena (a spatial skill chart). Another potential variation of the recommendation here would be to query the user community for product requirements at extended forecast timescales. In order to do this, the “user community” will need to be identified across Agencies. It was noted for this type of query that NOAA can’t go directly to private industry but could query other government agencies like the Departments of Agriculture and Energy (and could approach private industry indirectly through NSF and the National Academies, for example). Perhaps a user catalog could be developed to advertise what can be provided (seamless across timescales - not just a weather and climate catalog with two sections). This could essentially be a dynamic database for skill. In order to evaluate prediction skill at longer ranges, a comment was made to consider use of a scorecard, like that used for short-range forecasting, for longer-term forecasts. This may be a starting point for a more integrated physical approach.

Also during the Session 1 discussion, a need was expressed to demonstrate skill to users by specific skill on phenomena of interest to them (or consider use of other tools like a pattern based filter). The group also discussed upper bounds on model skill and confidence (confidence in range of errors, etc.). Does it make sense to compare uncalibrated models? Is this okay if using same data? It was noted that if you calibrate to one variable, you don’t generally account for effects on other variables. Perhaps information could be restored through other post-processing techniques. There is also the circular issue of what users want versus what forecasters can provide. Users need to be better educated about predictability.

Session 2 covered the topic of “Verifying Multi-model Coupled Systems - how do you trace errors and what metric do you use?”

While briefing the “NMME ISI Prediction Experiment,” Ben Kirtman gave 2013-14 NMME plans and discussed an overall skill assessment. He mentioned model upgrades (CCSM4, etc.) and data server work to increase data distribution. He also mentioned work to look closer at the subseasonal timescale – not real-time but to explore protocols for this (week 3-4 out to 12 months).

Huug van den Dool briefed “Issues/Challenges for MME Systems.” He showed stats for CFSv1 and CFSv2 along with other data (GFDL model, for ex). He indicated it is not clear whether NMME performs better than CFSv2 with perturbed physics (easier to set up, control, and maintain). However, there does appear to be a great urge to do the latter at NCEP, and so in the interim (which may be a long time), Huug indicated that the multi-model approach is a good distributed alternative that might work as long as partners (GFDL, NASA, NCAR, etc) are willing to go the extra mile. The attraction for the outlying partners to participate in NMME is the benefit resulting from exposure of their models to a particular real time application. During this talk, there was also discussion related to the value of shorter versus longer term hindcasts. A comment was made that a homogeneous data set for a 10-year data set is better than a non-homogeneous 30-year data set.

Bruce Cornuelle discussed the “Characterization of Model Errors through Comparison with Observations.” A question was posed of how you trace bias/drifts back to specific model developments (targeting development to address bias). Bruce showed examples to address vertical temperature biases in the ocean water column. Work is anticipated on ocean parameterization (sub-grid scale, tides, …). He also showed how other parameterizations or “nudging” impacting accuracy to forecasts of MJO for example.

In the Session 2 discussion period, several questions were posed. For example, in verifying multi-model coupled systems – how do you trace errors and what metric do you use? If the SST is too high, how do you discover that the cirrus clouds in the model are wrong? Do you look at the ocean mixed layer depth when trying to improve tropical cyclone genesis locations? A comment was made that we are getting good at developing statistical methods for measuring skill but not good at developing methods for measuring process. Additionally, we need to improve our skill in capturing relationships between different variables when tracing model error. There was debate over the importance of sample size on measuring processes (30 years vs. 10 years reanalysis – or do you go back to 1993 because that’s when satellite altimetry became available, for example). Tom Hamill commented that naïve use of a long data set isn’t good but he also said that it is not that hard to use trustworthy data over the longer period.

Further discussion on tracing model error considered the question of how error/bias is propagated through the ocean/air interface in coupled systems. What physical processes should you look into when you want to investigate errors/bias in forecasting a particular process? Although it would be nice to have a universal handbook covering this guidance, the guidance would not likely be the same for different models. And this will be particularly complex for a multi-model system. Depending on scale, you may also have to consider local influences. In terms of diagnostic tools, you may use adjunct sensitivity to sort out what part of model system might be the cause of a particular error mode. A comment was made that similar efforts were employed in some climate models (but again, there was discussion on that fact that tuning one variable or set of variables might have a positive impact in one area but a negative impact on another area(s). A broad set of automatically employed model diagnostics is a desired path. The group thought it might be helpful to consider past successful examples of tracing model errors. Cumulus momentum transport in GFDL model and positive impact on ENSO was mentioned as an example.

Resources are a limiting factor in analyzing the root cause for model improvement during development. Often a model is improved, but the modelers may say they don’t have the resources to analyze exactly what caused the improvement in the model. Several items may be modified/incorporated in a particular model update that results in an improvement. It may not be clear which of the modifications resulted in the improvement – and this may not be analyzed. Complementary skill of a given member in a multi-model ensemble can be demonstrated where a member has skill in a particular region over that of the other model(s). The underlying reason for the complementary skill (tracing to physical processes) is generally not known and is not investigated. A question was posed on whether models should be ranked in a particular order based on relative skill for a particular parameter or process (like permafrost, etc). Additionally, further consideration should be given to incorporating relative weighting of individual models in certain multi-models, situations, regions, etc. The group thought it would be useful to develop spatial skill maps for particular models at different lead times. Model bias by lat/lon and region is available for the Navy model (NAVGEM) for example, but it was not known whether an equivalent set for the ECMWF model is available for comparison. It was mentioned that NOAA NCEP has regional anomaly correlation performance by wave number ranges available.

Session 3 covered the topic of “Systematically Accounting for Model Inadequacy.”

Dan Copsey briefed “UK Model Development and Evaluation.” He provided the status and plans for the MOGREPS – the Met Office Global and Regional Ensemble Prediction System and for the GloSea5 seasonal forecasting system. Of note, plans are in place to institute coupling in MOGREPS to the NEMO ocean model and the CICE sea ice model. Dan briefed information on the testing of moving the UK 15-day forecasts to a coupled atmosphere-ocean-sea ice system.

The following are summary bullets from Dan Copsey’s brief:

- Some off line bias correction can be done for ensemble forecasts (e.g. wind thresholds)
- Coupled models outperform atmosphere only and ocean only forecasts in the following areas:
    - Improved forecasts of atmospheric air temperatures in the tropics
    - Improved forecasts of SSTs
    - Improved forecasts of tropical depressions
    - Correct lead-lag relationship between SSTs and atmospheric convection, improving the MJO and tropical precipitation near the Maritime Continent
- Coupled data assimilation shows potential for improving forecasts by removing initialization shock
- The UK Met Office seasonal forecasting system (GloSea5) has been upgraded recently with improved physics, horizontal and vertical resolution. These have improved:
    - Predictability of the NAO, due to reduced mean state biases (no cold SST bias in the North Atlantic) and improved teleconnections (e.g. ENSO via sudden stratospheric warmings)

Gabe Vecchi briefed “GFDL Seasonal to Decadal Predictions: Challenges and Issues.” He stated that more and deeper ocean profiles are available but some issues exist with handling the data. Initialization leads to skill predictions of the sub-polar gyre. He hypothesized that enhanced resolution (particularly for the atmosphere and land) will lead to improved simulation and prediction of climate. The goal of the improvement would be to build a seasonal to decadal forecasting system to yield improved forecasts of large-scale climate and to enable forecasts of regional climate and extremes.

Gabe Vecchi stated the need for creative methodologies that account for changes in observing systems. Strong supportive efforts are needed for successful initialized prediction (in coupled model development, assimilation develop, sustained observing systems, and in analysis methodology). Initialized predictions are yielding exciting results on seasonal to decadal timescales. Encouraging results are being shown in seasonal to decadal prediction with for example, CM2.5 FLOR towards seamless seasonal to centennial prediction/projection of regional and extremes.

The following are summary bullets from Gabe Vecchi’s brief:

- Successful initialized prediction requires strong efforts in:
    - Coupled model development
    - Assimilation development
    - Sustained observing systems
    - Analysis methodology
- Initialized predictions yielding exciting results on seasonal to decadal timescales
    - Dealing with observing system inhomogeneity remains a challenge
    - Model bias impacts skill through drift/shock, accentuating impacts of observing system changes and complicating assimilation analysis
- Initialized predictions test models in ways “free running” simulations do not: valuable
- High-resolution seasonal to decadal predictions
    - CM2.5-FLOR: towards seamless seasonal to centennial prediction/projection of regional and extremes (50km/CM2.5 atmosphere/land - Delworth et al. 2012; 1°/CM2.1 ocean/ice – Delworth et al. 2006)
    - 4500+ years of retrospective forecasts: encouraging results on precipitation, continental temperature and tropical cyclones

Pat Hogan briefed the “Gulf of Mexico Pilot Prediction Project” and provided preliminary global ensemble results. The overall goal of the project is to implement the capability to provide a long-range forecast (60 days). The methodology and value of using an ensemble modeling approach in the project was presented. Reasonable qualitative skill in forecasts of Gulf of Mexico sea surface temperature and eddy evolution were shown through eight weeks in preliminary results. Prediction of loop current eddies is of particular importance to drill rig crews because equipment (drill stems) may need to be pulled out at a given ocean current increase cut-off.

During the Session 3 discussion, the group considered whether a sufficient understanding exists of the ways different approaches to accounting for model inadequacy in ensembles impact results in order to develop ensemble prediction systems with desirable properties. Different model resolutions drive different parameterizations and forcing input. Limited resources make it difficult to fully explore the impact of different structure (different resolution or parameterization impact, etc.) and payoff in conducting the analyses may be negligible. The impact of insufficient coverage of model uncertainty due to poor coverage of spread was mentioned as a negative issue for users. There is the question of whether model developers are doing enough to tease out the full uncertainty in models. For a multi-model ensemble, a desire was stated for the development of automated tools that provide information on the impact of particular initial conditions on results (when considering perturbation of initial conditions between different ensemble members). Tom Hamill noted that we have inherited deterministic parameterizations for ensemble use. Re-thinking/reconceptualization of this issue is needed. Perhaps the use of stochastic physics is an answer here. A comment was made that the scientific way to treat this issue is at the process level. Parameterizations and stochastic physics, however, only address one part of uncertainty. Barbara Brown also mentioned post-processing (Bayseian Model Averaging (BMA), etc.) as a way of addressing model differences and inadequacies. Stan Benjamin asked whether different members of NMME could conduct equivalent experiments to compare results from different models – for example, apply a common stochastic physics approach to different member models and compare results.

The group expressed a general dissatisfaction with ad hoc approaches for studying model uncertainty but understood the difficulties associated with more deliberate approaches. For example, there aren’t enough interchangeable parts between models. Combining models helps you tease out the full PDF but you can’t attribute specific skills to specific models or aspects of these models in many cases. Automated tools to attribute impacts (map skill indices to model pedigree) on parts of PDF to specific models (or their aspects) or model changes, etc. would be useful. Use of a single column model concept across modeling groups could apply here. The question was asked of whether there are experiments that could show linkage of a multi-model ensemble to specific parts of a PDF (show more detail on where complementary skill comes from). Liz Satterfield has an upcoming project to investigate how multi-model ensembles yield benefits (attribution of multi-model ensemble skill to particular results). There is also the question of whether tracking uncertainty is tractable once you begin to couple the atmosphere/ocean/land/ice, etc. The problem certainly becomes more complicated.

Session 4 of the workshop covered the topic of "Post-Processing and Visualization on ISI Timescales."

Liz Satterfield briefed "Heteroscedastic Ensemble Post-Processing." She introduced her topic by saying:

- Probabilistic predictions provide information that can increase the socio-economic value of weather and climate forecasts
- Ensemble based probabilistic forecasts of events are typically based on the relative frequencies of events in the ensemble
- Accurate predictions of forecast error variances are vital to probabilistic prediction

Accurate forecast error variances are needed. As sample size increases, the distribution of variance sharpens. Liz Satterfield's efforts considered effective ensemble size. Ensemble size is not necessarily the sample size. An effective post-processing scheme will account for stochastic fluctuations of the ensemble variance (ability of the ensemble to track fluctuations in the variance). Uncertainty is reduced because ensemble subspace captures error. Liz Satterfield concluded:

- A new post-processing scheme was presented which accounts for a distribution of possible variances given an imperfect ensemble predictions
- Bishop et al 2013 introduce a new diagnostic, the effective ensemble size, which measures the ability of the ensemble to track fluctuations in actual error variance
- Application to synthetically generated data and 500hPa forecasts of virtual temperature from the operational FNMOC ensemble demonstrate that accounting for the variable nature of forecast error variances leads to improved probabilistic skill scores

Josh Hacker briefed "Predictability and Calibration Beyond the Medium Range." He presented some thoughts on presenting forecasts to users. He stated that giving a user only the calibrated forecasts eliminates/minimizes user ability to drive model improvement. From users, we desire to know what aspect of a model forecast is important for their decision process. He also expressed a need to put actual forecasts in front of people. Let them interact with the data. Track where what data they use and how they use it. This approach indicates both trust and utility.

Jin Huang gave a presentation on "Climate Prediction Center (CPC) Products and Decision Support." She discussed:

- CPC overview and product suites
- Forecast format, tools, post-processing, and verifications
- Examples of decision support products
- CPC future plans

In general Jin Huang summarized that:

- CPC develops, verifies and disseminates official climate forecast products for ISI timescales
- CPC post-processing includes:
    - Objective consolidation of dynamical and empirical tools

    - Subjective decisions for the official products
- Forecast format, tools, procedures and verification metrics vary with timescales and products
- CPC plans to introduce new products (and stop underutilized products) in collaboration with the research community, the weather community and other agencies
    - Climate Test Bed is a mechanism for R2O and O2R

David Margolin provided a briefing on "A Trader's Perspective on Why Weather Matters for Energy." He discussed a brief background on natural gas trading and weather's role in natural gas trading. He also presented a trader case study to include the daily challenges of being a natural gas trader (how traders leverage weather information, along with other market information, to evaluate trades). He summarized that weather remains a core component of any fundamental trading strategy. From his perspective in moving forward:

- Sophisticated analytics and cutting-edge research are valued more than in the past, but in many cases they cannot be leveraged by the end user
    - Probabilistic forecasts (Full Distribution)
    - Skill metrics targeted for the user (i.e., extreme events are important in Energy). Hit -Rates are more tangible than Ignorance for most end users.
    - Model Diagnostics that the operational forecaster can leverage real-time
    - User-Defined weather forecasts (i.e., the natural gas industry is most interested in the NYC-ORD-ATL corridor)

Prices for natural gas react to weather because of demand/usage. The trading industry sees more sophisticated analytics and cutting edge research as valuable but may not be able to leverage these. For example, end users may not relate to certain metrics but generally do like hit rates. Industry forecasters would like more model diagnostics (for example, more info on extent of cold bias in a particular model at different lead times). Forecasts out to 15 days are utilized (11-15 day forecasts included). Week 3 and 4 forecasts are becoming more relevant. Traders like to see seasonal forecasts but don't really use them. Delivery timing is also important to timely (and sometimes irrational) trading decisions. If all users don't operate off of the same schedule of product delivery, users that receive information earliest will benefit if the forecast information is accurate. David Margolin said the industry is ready for more probabilistic products (such as depictions of bins of probability for particular forecast values for decision-making).

Several questions were posed for the Session 4 discussion. How do we adjust/calibrate 30-day to annual predictions to enhance predictive skill and customer value? Do we need to define customers to proceed? One approach to addressing these questions is to provide data in easily accessible format/locations that make it easy for customers to pull data and format in their own tailored/desired fashion. The group recognized that different classes of users exist and interaction with the different groups to define needs may need to be tailored. The group agreed that better tools are needed for any end user to better mine data.

On the first question of calibration, one opinion was stated that data provided to the general public should not be calibrated. But perhaps we could interact with users desiring to calibrate (to understand their needs/issues). A recommendation was to provide the supportive databases to allow for intelligent calibration efforts. Barbara Brown mentioned NRC report recommendation that said NOAA should make all data available for applications (would include end user calibration of model information). The bottom line of this discussion was that customer interaction is absolutely necessary while recognizing there are restrictions on tailoring of data/information for certain customers. Other interactions with user communities can be conducted through use of buffering activities or organizations like AMS or National Academies, etc. where sensitivities exist such as direct interactions between NOAA and private entities.

Session 5 focused on “Constructing a Framework for Formulation and Inter-Comparison of Prediction Systems.” Leading off the session Barbara Brown provided a briefing on “Constructing a Developmental Test Center (DTC) for ISI.” Barbara stated that global modeling is likely in DTC’s future. The general focus of the DTC is to support sharing, testing, and evaluation of research and operational numerical weather prediction systems. The mesoscale model evaluation testbed may be a model of how the ISI DTC might be established. The DTC would have an O2R part – many options here but could make operational codes available for researchers to work directly with it –and- provide information about operational needs to researchers. The DTC would also have an R2O part, including testing of new capabilities and ingesting new research capability into operations, for example. This DTC would need to dovetail with the CTB plans (not clear at this point).

Jin Huang provided a briefing on the “NCEP Climate Test Bed (CTB) as a Development Test Center for ISI – CTB plans.” She briefed CTB’s mission and organization structure, current priorities and activities, and evaluation metrics for CTB R2O transition. She also indicated:

- CPC develops, verifies and disseminates official climate forecast products for ISI timescales
- CPC operational forecast system and NMME provide an platform for model inter-comparison and evaluation in an operational setting
- CTB is a NCEP test bed dedicated for testing and improving ISI climate forecast and products (but needs dedicated FTEs (as other testbeds, e.g. JCSDA, DTC) to better serve broader stakeholders and users

During the Session 5 discussion, there was a question posed on whether ESRL would want to pump developmental model data into a central location/testbed where common diagnostics would be applied or would they prefer to run their own diagnostics? A comparison was made between the Hurricane Forecast Improvement Project (HFIP) and the ISI global prediction effort in this context. HFIP has a discrete project goal and quantitative metrics. The challenge for ISI global prediction is the wide span of foci/targets that can be chosen. A perspective was offered that Phase II of NMME should focus participation and maybe Phase III, in line with ESPC, will focus target(s) further. Dan Eleuterio expressed a desire to establish a test bed with a set of models that meet an established set of criteria. Some data manipulation tools could also be available.

Workshop Recommendations:

1. Ensemble skill metrics have been shown to be dependent on the selected target parameter. Identify user-driven metrics that will stress the coupled model parameter space and the ISI temporal space.

- These can be developed formally through user workshops/interaction
- These can be a survey of NMME/CPC participants about their customer interactions

Related recommendations:

- Develop a user "scorecard" of five or six targets
    - Develop a table of user requirements by area, parameter, threshold, time
    - Sponsor a workshop for ISI users
    - Put forecasts in front of users and let them interact
    - Need to state explicitly whether we will or will not engage "users" defined as atmos/ocean/ice forecasters that support decision makers

2. The scientific hypothesis in support of the multi-model approach is that the diversity of models on average resolves the PDF better than an individual model. One test of this hypothesis is to diagnostically evaluate the complementary information provided in multi-model ensembles. For example, given the NMME retrospective data set, it is possible to rigorously evaluate the complementary skill provide by

each of the forecast systems. The recommendation is to encourage/support/enable the analysis of multi-model experiments (CMIP5, TIGGE, NMME , …) to determine the complementary information/skill.

Related recommendations:

- Investigate improved means of providing data in easily accessible format/locations that make it easy for customers to pull data and format in their own tailored/desired fashion (could include providing the supportive databases to allow for intelligent calibration efforts)
- Have data in digital interactive data base so users can set own thresholds, etc., don't calibrate or post process, but assist user if necessary
- Develop metrics appropriate for each time/space scale
- Use reforecasts to determine skill at larger spatial and timescales
- Develop a global map of forecast skill or predictability
- Define comprehensive diagnostics for model improvement
- Studies to identify the degree to which multi-model ensembles provide improved forecasts due to independence vs. due to offsetting errors
- Metrics need to show geographic or temporal variability; single-number boil-downs have their uses but are inadequate to customers. We need products to show forecast uncertainty, in addition to quality. For knowledgeable users, a way to show clumping of model solutions would be illuminating.

3. ESPC should support the development of coupled model data assimilation. Methods to trace errors in a complex coupled ensemble system.

4. We should reconceptualize how we parameterize models for ensemble prediction at ISI timescales. Do we adopt stochastic parameterizations? Does this change the number of ensemble members required?

5. Participate in S2S community to increase seasonal data availability, archiving, and data tools.

6. Define and provide data set protocols, tools, repository, support and funding mechanisms to involve the broader community in the NMME and improving ISI prediction and prediction systems.

7. Longer range goal is to improve multi-model ensemble (NMME, IMME and other) guidance for seasonal forecasting.

8. Focus an upcoming NCAR summer colloquium (Advanced Studies Program) on diagnostics for coupled and S2S models and NMME Phase II.